3В

УДК 004.934

Н.Б. Васильєва
Міжнародний науково-навчальний центр інформаційних технологій та систем,
м. Київ, Україна
ninel@uasoiro.org.ua

Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання злитого мовлення

В статті представлені результати поточних досліджень багаторівневого багатозначного підходу до автоматичного розуміння мовлення, який призначений для мов з розвинутою словозміною та з відносно вільним порядком слів. Ідея використовувати мовленнєві образи на рівні частин слова для пофонемного розпізнавання є продуктивною, оскільки зростання обсягу лексикону практично не призводить до збільшення множини частин слова. Дослідження зосереджено на рівні пофонемного розпізнавання з метою подальшого переходу на лексичний рівень. Щоб зменшити показник помилково розпізнаних фонем, будується граматика декодера на основі вільного слідування частин слова. Подаються способи побудови множини (алфавіту) складів за текстовим корпусом. Описуються підходи до формування навчальної та контрольних вибірок для пофонемного розпізнавання злитого мовлення. Для експериментальних досліджень проведено роботу зі створення корпусу опорного диктора. Порівнюється надійність пофонемного розпізнавання з використанням граматик на основі складів і фонем. Обговорюються придатність отриманих результатів для використання в лексичному рівні, проблеми та майбутні дослідження.

Вступ

Загальноприйняті системи пофонемного розпізнавання оперують алфавітом фонем (контекстно залежних або контекстно незалежних), з яких складаються мовленнєві образи слів. Вже на слова накладаються обмеження їх слідування шляхом побудови граматик або лінгвістичної моделі (ЛМ). При збагаченні лексики зростають обсяги робочого словника, суттєво ускладнюються граматика або ЛМ, а це призводить до зменшення продуктивності системи розпізнавання.

Якщо використовувати замість слів мовленнєві образи складів або морфем, то збагачення лексики не призведе до помітного зростання робочих словників та ускладнення граматики чи ЛМ.

В цьому випадку найбільшою перепоною є перехід від послідовностей складів (морфем) до послідовностей слів, оскільки помилка розпізнавання складу або морфеми може спричиняти ситуацію, коли їх послідовностям неможливо співставити слово. Також сама процедура переходу до послідовностей слів є неоднозначною і малодослідженою.

В попередній роботі [1] описувалося проведення поширення багаторівневої багатозначної моделі автоматичного розпізнавання злитого мовлення на випадок поскладового розпізнавання. Для проведення експериментальних досліджень застосовувався багатодикторний мовленнєвий корпус. Навчальна вибірка, сформована на основі цього корпусу, складалася з відносно невеликої кількості ізольованих слів. Використовувався словник лише на 4 000 слів за наявних близько 20 тис. реалізацій цих слів, вимовлених 70 дикторами. Результати проведених експериментів могли бути порівняні лише між собою, оскільки ця база ніким більше не використовувалася, більше того, на сьогоднішній день не існує доступного і придатного для наукових досліджень україномовного мовленнєвого корпусу.

З огляду на це, було вирішено сформувати свій однодикторний мовленнєвий корпус, в якому би спостерігалося все розмаїття звуків української мови. На основі цього корпусу планувалося проводити експерименти, порівнювати результати розпізнавання як окремо вимовлюваних слів, так і злитого мовлення для різних мовленнєвих образів, а також порівнювати результати розпізнавання з результатами розпізнавання на багатодикторному мовленнєвому корпусі.

Метою даної статті є підвищення пофонемної надійності розпізнавання злитого мовлення, що створить передумови реалізації ефективних алгоритмів переходу від послідовностей складів (морфем, фонем) до слів [2].

У другому розділі статті описується формування навчальної вибірки, у третьому – формування контрольних вибірок. Четвертий розділ присвячено експериментальним дослідженням пофонемного та поскладового розпізнавання. Наприкінці обговорюються результати та подальші дослідження.

Формування тексту та акустичної бази навчальної вибірки

Для проведення навчання і тестування розпізнавання, а в подальшому синтезу, необхідно мати широку експериментальну базу, куди входять:

– однодикторні або багатодикторні навчальні і контрольні вибірки (НВ і КВ) для дослідження індивідуалізованого і кооперативного розпізнавання;

– текстовий корпус для формування алфавітів мовленнєвих образів і текстів для запису мовленнєвого корпусу.

Формування мовленнєвої бази даних і знань потребує великих затрат часу, а також детальної підготовки тексту НВ. Для отримання НВ, що містить якомога широкий спектр різних фонем-трифонів, використовувалися тексти, що знаходяться у вільному доступі в Інтернеті. В основному це: художні твори українських авторів, публіцистичні твори, новини, історичні довідки. Виключалися віршовані твори, оскільки вірші, на відміну від прози, читаються з особливостями, не властивими повсякденному мовленню (інший інтонаційний контур, нестандартне наголошення складів тощо). Для НВ з ізольованих слів використані частотний словник української мови та словник УМІФ [3].

У процесі формування тексту НВ було здійснено такі заходи:

– попередня обробка текстів (видалення приміток, номерів розділів, заміна скорочень тощо);

– виділення окремого речення в рядок;

– перетворення орфографічного тексту в фонемний;

– складання статистики по кожному із обраних початкових корпусів (підраховується скільки разів зустрівся той чи інший мовленнєвий образ та найкоротше речення, де він зустрівся);

– оброблення отриманого результату через «Жадібний» алгоритм (ЖА) [4].

У вибрані таким чином речення потрапляють ті, які містять нову фонему-трифон і які є найкоротшими із можливих, що розглядаються. Таким чином, отримуємо суттєве скорочення тексту НВ, не втрачаючи фонемного розмаїття. Деякі статистичні дані за різними джерелами (текстовий корпус, словник УМІФ та частотний словник) в отриманих НВ наведені в табл. 1.

Таблиця 1 – Кількість фонем-трифонів, зареєстрованих лише один раз

Назва початкового текстового корпусу, з якого вибиралася НВ	Кількість фонем-трифонів, які зустрічаються лише один раз		Загальна кількість фонем-трифонів у кожній з НВ
	Початковий текст	НВ	
Текстовий корпус	5072	23972	51442
Словник УМІФ	1720	17372	27772
Частотний словник	2273	10798	18337

Акустичні моделі формувалися на основі контекстно незалежних фонем, оскільки обсяг їх алфавіту невеликий, а отже для статистичних оцінок потрібна менша база акустичних сигналів, ніж для складів або фонем-трифонів, яких більше в тисячі разів і топологія їх акустичних моделей потребує додаткових досліджень.

На першому етапі (етапі обробки текстового корпусу і формування НВ) розглядалися фонеми-трифони як мовленнєві образи, оскільки вони мають регулярну структуру і дають змогу моделювати фонемне розмаїття шляхом врахування правого і лівого контексту [5]. Структурно фонема-трифон включає три символи на відміну від складів, які можуть містити різну кількість фонем, а отже, символів: від однієї до шести (або до восьми символів при написанні з позначенням пом'якшень та наголошених голосних) у складах, які діляться за правилами складоподілу, і до п'яти (та семи символів при написані) у відкритих складах.

Застосування ЖА приводить, зокрема, до того, що автоматично видаляються дуже схожі речення, включно з повторами (одне речення може містити декілька унікальних фонем-трифонів і таким чином потрапить у навчальну вибірку декілька разів).

У табл. 2 показана статистика фонемів-трифон у НВ, оптимізація при роботі ЖА.

Наступним етапом створення НВ є запис мовлення за сформованим текстом навчальної вибірки. Під час запису проводиться апробація отриманих результатів на зручність читання, перевірка транскрипції, виявлення помилок, які не виявляються автоматично і які заважать нормальній вимові диктора тощо.

При опрацюванні текстів не було змоги врахувати такі пізніше виявлені проблеми:

– помилково написані фрази-речення (написані правильно орфографічно, але позбавлені семантики);

– друкарські помилки;

– візуальна схожість літер у кирилиці та латиниці: а, о, е, у, і, р, с, х, Е, Х, Н, В, А, О, Р, М, Т;

– літера замість цифри (в основному це стосується римських позначень цифри) та навпаки;

– скорочення типу 1-ї, 1-го, 1-е, 1-й, 1-м, 1-у, 1-ї, 1-у;

– написання та вимова слів, означених цифрами (календарні дати, дробові числа та інше);

– ізольована літера, наприклад, в кінці фрази, після якої стоять три крапки, або перед цифрою тощо.

Таблиця 2 – Порівняння кількості елементів у початковому корпусі та тексті НВ

Початковий текстовий корпус, з якого вибиралася НВ	Загальна кількість речень (слів для словників) до роботи ЖА	Загальна кількість речень (слів) після роботи ЖА	Загальна кількість реалізацій фонем-трифонів до роботи ЖА	Загальна кількість реалізацій фонем-трифонів після роботи ЖА	Алфавіт фонем-трифонів
Текстовий корпус (709 файлів; ~ 50 МБ)	815 995	18 019	41 179 780	102 0356	51 442
Словник УМІФ	1 874 744	13 706	23 734 396	12 0194	27 772
Частотний словник	137 634	8 207	1 487 679	71 019	18 337

У процесі запису спостерігалися такі фізіологічні та психолінгвістичні явища:

– втома голосового тракту;

– зміна голосу в різних життєвих ситуаціях (захворювання, хвилювання, час доби тощо);

– специфіка вимови деяких словосполучень та словоформ (редукція та асиміляція за глухістю і дзвінкістю приголосних звуків).

При обробці словників також була проведена попередня робота перед записом, яка полягала в тому, що було вилучено слова, що містять однакові фонеми-трифони, оскільки ми маємо два словника, які мають певну кількість однакових фонем-трифонів, тобто дублювання.

Кожен із словників має унікальний алфавіт фонем-трифонів. Наприклад, словник УМІФ має таких фонем-трифонів 12 327, а частотний словник – 2 916. В двох словниках одночасно виявлено 15 431 однакових фонем-трифонів.

Запис НВ і наступних вибірок проводився за допомогою модуля Sigrs [6] на звуковій карті Creative Audigy2 ZS гарнітурою SteelSeries 5H v2. Отримано близько 36 годин запису злитого мовлення. Обсяг словника НВ – 47 621 слів. Загальна кількість реалізацій слів у НВ – 184 910.

Обсяг словника НВ окремих слів складав 12 870 слів, близько 12 годин запису.

Формування контрольної вибірки

Було вирішено сформувати свій однодикторний мовленнєвий корпус, в якому би спостерігалося практично все розмаїття звуків української мови. На основі цього корпусу планувалося проводити експерименти, порівнювати результати розпізнавання як окремо вимовлюваних слів, так і злитого мовлення для різних мовленнєвих образів, а також порівнювати результати розпізнавання з результатами розпізнавання на багатодикторному мовленнєвому корпусі.

Для перевірки запропонованих мовленнєвих образів, тобто фонем, відкритих складів та складів, поділених за правилами складоподілу, було сформовано тексти контрольної вибірки (КВ) злитого мовлення та проведений її запис.

Вирішено було провести тестування на двох КВ, сформованих різними способами.

Частотна КВ

Перший спосіб вибору КВ вирішено базувався на тому, щоб перевірити на розпізнавання часто вживані слова, речення, фрази, тобто сформувати КВ за частотністю фонем-трифонів.

Етапи отримання частотної КВ:

– з початкового текстового корпусу вилучається текст НВ;

– із залишеного текстового корпусу формується статистика по фонемам-трифонам та найкоротші речення, в яких вони зустрілися;

– з цих речень береться деяка кількість перших речень (в нашому випадку – 3 000);

– вилучаються речення, які повторюються.

Запис проводився в тих самих умовах, що і запис НВ.

Отримана КВ має 3,6 години запису. Обсяг словника складає 3 225 слів. Загальна кількість реалізацій слів – 8 987.

КВ, сформована випадковим чином

Другий спосіб – сформувати КВ випадковим чином, з тих самих текстів, з яких вибирався текст НВ, але з забороною вибору тих речень, що ввійшли до НВ.

Етапи отримання випадкової КВ:

– з початкового текстового корпусу вилучається текст НВ;

– із залишеного текстового корпусу випадковим чином береться деяка кількість речень (в нашому випадку – 2 000);

– вилучаються речення, які повторюються.

Запис проводився в тих самих умовах, що і запис НВ.

Отримана КВ має 4,3 години запису. Обсяг словника складає 10 013 слів. Загальна кількість реалізацій слів – 22864.

КВ «Вікіпедія»

Цю КВ запропоновано вибрати з текстів, які не використовувалися ні для вибору попередньої КВ, ні для НВ. Для цього з сайту української Вікіпедії [7] випадковим чином вибрано тексти.

Етапи отримання КВ «Вікіпедія»:

– з текстів сайту Вікіпедія вилучено речення, що зустрілися НВ та КВ випадкової;

– із отриманого тексту випадковим чином вибираємо 1000 речень;

– вилучаються речення, які повторюються;

– додано 200 послідовних речень з однієї випадково обраної статті.

Запис проводився в тих самих умовах, що і запис НВ.

Отримана КВ має 3,0 години запису. Обсяг словника складає 7 330 слів. Загальна кількість реалізацій слів – 16 073.

Експериментальне розпізнавання та порівняння отриманих результатів

Було проведено оцінювання параметрів акустичних моделей з використанням програмного інструментарію HTK [8] для кожного з 57 монофонів і двох фонем-пауз. Отримані моделі фонем мають кожна три стани і від 4 до 36 сумішей нормальних законів в залежності від акустичної мінливості та частотності.

При розпізнаванні допускалася вільна граматика слідування фонемних образів як для фонем, так і для складів. Тільки для відкритих складів було накладене обмеження: склади, які не мають голосної, стоять перед паузою.

Процедура розпізнавання проводилася за допомогою програмних комплексів HTK та Julius [9] на трьох контрольних вибірках: частотній, випадковій та «Вікіпедія». Одиницею розпізнавання взято три фонемні образи: фонеми (кількість одиниць 59), відкриті склади (кількість одиниць 17 270), та склади, поділені за правилами складоподілу (кількість одиниць 10 200).

Відповіді розпізнавання зводилися до пофонемного вигляду з метою подальшої оцінки надійності порівняно до еталонного фонемного тексту. У табл. 3 наводиться фонемна похибка (англійською, PER – Phoneme Error Rate) для описаних вище контрольних вибірок. Зауважимо, що до алфавіту фонем входять, зокрема, наголошені та ненаголошені голосні фонеми. На вимові плутання голосної та приголосної може призвести до спотворення змісту. Втім, на письмі наголос зазвичай опускається. Виходячи з цих міркувань, в результатах розпізнавання також подається похибка без урахування наголосу, що дало значно меншу оцінку PER.

І хоч для ряду експериментів про результат говорити зарано, очевидним є факт залежності похибки від методу формування КВ. Так тексти з вибірки «Вікіпедія» не входили до початкового текстового корпусу, а отже ця вибірка містить певну кількість

фонем-трифонів, відсутніх у НВ. «Випадкова» КВ відповідає загальній статистичній картині, тому слід орієнтуватися на показники надійності саме для цієї вибірки. Також слід зазначити, що довжина речення для «Частотної» КВ) складає в середньому 3,2 слова, тоді як у «Випадковій» КВ середня кількість слів у реченні становить 10,5, тобто майже як в НВ. Подальші дослідження НВ ізольованих слів для навчання можуть пояснити ці результати.

Таблиця 3 – Показники помилково розпізнаних фонем (%) для злитого мовлення на основі різних мовленнєвих образів з використанням інструментарію HTK та Julius

КВ	Фонема		Відкритий склад		Склад за правилами складоподілу	
	HTK	Julius	HTK	Julius	HTK	Julius
«Випадкова» КВ	28,86	29,11	24,92	24,46	24,54	24,03
«Випадкова» КВ (без наголосів)	21,39	22,28	17,68	17,47	17,29	17,01
КВ «Вікіпедія»	31,93	35,48	28,01	30,17	28,18	31,08
КВ «Вікіпедія» (без наголосів)	24,72	23,19	28,81	20,81	21,00	22,37
«Частотна» КВ	36,6	-	37,75	-	-	-
«Частотна» КВ (без наголосів)	26,1	-	27,95	-	-	-

Висновки

Порівняно з попередніми дослідженнями [1] фонемна похибка розпізнавання для злитого мовлення в окремих випадках зменшилась більш ніж на половину. Це зумовлено вдосконаленням навчальної вибірки для оцінки параметрів акустичних моделей та врахуванням індивідуальних особливостей вимови диктора. Поки що з результатів чітко не простежується, який вид складоподілу кращий.

Запропонований спосіб формування навчальної вибірки дає змогу широко охопити фонетичне розмаїття мови, використовуючи близько 2% речень від усіх спостережуваних.

У наведених експериментах допускалася вільна граматика слідування частин слів. Планується застосувати статистичні лінгвістичні моделі для частин слів, що має привести до зменшення помилок розпізнавання.

Планується задіяти в експериментальні дослідження частину навчальної вибірки з окремо вимовлюваними словами. Залишаються недослідженими вплив багатьох параметрів декодера на надійність та швидкість. Зокрема, розроблятимуться підходи до зменшення алфавіту складів, що повинно прискорити розпізнавання.

Подальші дослідження покажуть, наскільки досягнутого рівня надійності достатньо для переходу від послідовностей фонем (з супровідною оцінкою акустичних параметрів) до послідовностей слів на лексичному рівні багаторівневої багатозначної системи розуміння мовлення.

Література

1. Vasylieva N. Modelyuvannya bahatorivnevoho poskladovoho rozpiznavannya movlennyevoho syhnalu / N. Vasylieva, M. Sazhok // Artificial Intellect –2008.– № 3. – S. 801-808.
2. Sazhok M.. Generative Model for Decoding a Phoneme Recognizer Output. / Mykola Sazhok // Proc. of the 8th International Conference «Text, Speech and Dialogue», TSD'2005. – Karlovy Vary, 2005. – P. 288-293.
3. Shyrokov V., Monako V. Organization of national lexicographic framework resources./ V. Shyrokov, V. Monako // – Movoznavstvo. – 2001.– № 5.
4. Goncharov E. Behavior of probabilistic greedy algorithm for stage location problem. / Goncharov E., Kochetov Yu. // Sampling analysis and operations research.– 1999. – Vol. 6, № 1. – S. 12-32.
5. Young S.J. HTK Book.– Young S.J. version 3.1. – Cambridge University, 2002.
6. Lee T. Julius – an open source real-time large vocabulary recognition engine./ Lee, T. Kawahara and K. Shikano // Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001. – P. 1691-1694.

Н.Б. Васильева

Использование грамматик свободного порядка следования фонем и слогов для пофонемного распознавания слитной речи

В работе представлены результаты текущих исследований в рамках подхода многоуровневого многозначного автоматического понимания речи, предназначенного для языков с развитым словоизменением и с относительно свободным порядком слов, поскольку рост объема лексикона практически не приводит к увеличению множества частей слова. Предлагаются способы построения множества (алфавита) слогов по текстовому корпусу. Описываются подходы к формированию обучающей выборки и контрольных выборок для фонемного распознавания слитной речи. Сравнивается надежность фонемного распознавания с использованием грамматик на основе слогов и фонем. Пригодность результатов к использованию в качестве входных данных лексического уровня является предметом будущих исследований.

N.B. Vasylieva

Free Phoneme and Syllable Order Grammar Application for Continuous Speech Phoneme-by-Phoneme Recognition

The paper presents advances in a multi-level automatic speech understanding approach that is initially developed for highly inflective languages with relatively free word order since word lexicon growth leads to practically no new sub-word items. The ways to select a set of sub-word units like syllables are considered. The proposed procedure to select a set of sentences containing all phoneme-triphones allowed for creation the text for training corpus. Three control sets were formed by different ways. The recognition accuracy has been compared to free phoneme order grammar. The results show the promising input for the next lexical level of the multi-level automatic speech understanding system.

Стаття надійшла до редакції 22.06.2011.